tMANUAL

5/31/23

Table of contents

Preface

tMANUAL is a guide to using tMAVEN.

For more information visit:

- [Website] https://gonzalezbiophysicslab.github.io/tmaven/
- [Github] https://github.com/GonzalezBiophysicsLab/tmaven

1 1. Installing

tMAVEN is a free open-source software available to anyone looking to comparatively analyze single molecule time-series data. The easiest way to Install tMAVEN is to navigate to https://github.com/GonzalezBiophysicsLab/tmaven, go to releases and install the latest .dmg for Mac or .msi for Windows.

Otherwise, users can also do a manual installation (not recommended) also oulined at https://gonzalezbiophysicslab.github.io/tmaven/install.html. There are a few softwares necessary for installation:

- Python, which is available for download at many sites, we recommend miniconda which may be installed at https://docs.conda.io/en/latest/miniconda.html

If this software is already set up, a user must do the following in sequence:

- navigate to the tMAVEN repository at https://github.com/GonzalezBiophysicsLab/tmaven, and clone the repository.
- open terminal navigate to the tMAVEN folder and enter the command `pip install -e .` and select yes/enter to any subsequent requests.

2 2. Navigating

tMAVEN can support ASCII, SMD [Greenfeld et al. (2015) ; HDF5 based], raw HDF5, and SPARTAN files. Load data by hitting File/**Load**, select file type, and then the file itself. Note that class, pre/post *etc.* files may be uploaded separately. The default upload settings are for SMD files, so if using ASCII files, the upload setting must be chosen through the menus File/Load/Text files (ASCII/UTF)/Raw/[user's format]. In this section, N refers to number of traces, T to number of time points and C to number of colors. Collation axis (0 or 1) refers to how the text separates individual traces
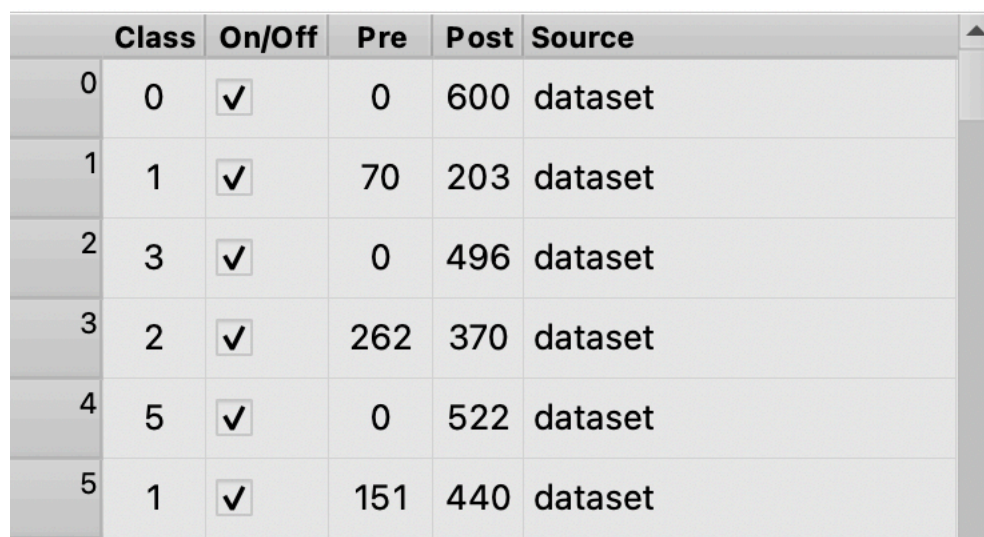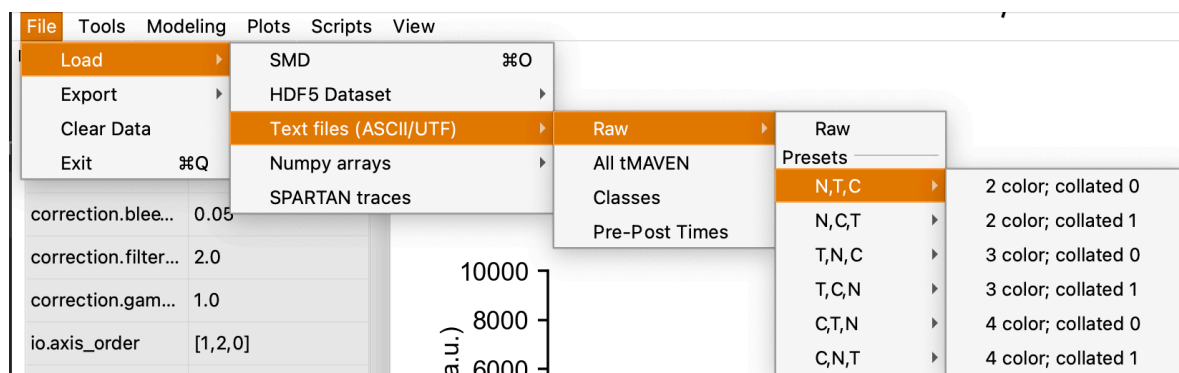

Note that any preferences using True/False are case sensitive. There is also an **HDF5 viewer* tool under "View'' that the user may find useful.

3 The Molecule Table

Navigate to the molecule table View/**Molecule Table** or hit command T to bring up the molecule table. Notice the columns:

- Class: the class of the trace, change a trace's class by clicking its row in the molecule table and inputing a number, or simply by clicking around the plot and keying a number.
- On/Off: Whether the trace is turned on. By toggling a trace one can easily choose which traces are included in modeling and plotting features, see section 4. It will also toggle whether the trace is subject to actions like save.
- Pre/Post: The range of time each cut trace covers
- Source: If modeling from different datasets, the origin of each trace can be seen here. Note that the class can be assigned separately from the source.

	Class	On/Off	Pre	Post	Source
0	0	✓	0	600	dataset
1	1	✓	70	203	dataset
2	3	✓	0	496	dataset
3	2	✓	262	370	dataset
4	5	✓	0	522	dataset
5	1	✓	151	440	dataset

Figure 3.1: Molecule Table Sample

Classifying traces is one of the most powerful aspects of tMAVEN.

4 Using the UI and Cutting a Trace

Navigate between traces by scrolling while graph window is selected or by using the scroll underneath the graphs (see Figure). Notice trace number and class of the graph displayed in the bottom right corner. Aspects of the graph can be adjusted in the graph settings and figure options (see Figure), or in preferences. Most of these settings are plot.[aspect] and include settings for the axes, subdivision ticks, time scale, and more. The user can always use the View/**reset GUI** command to set all these settings to default.
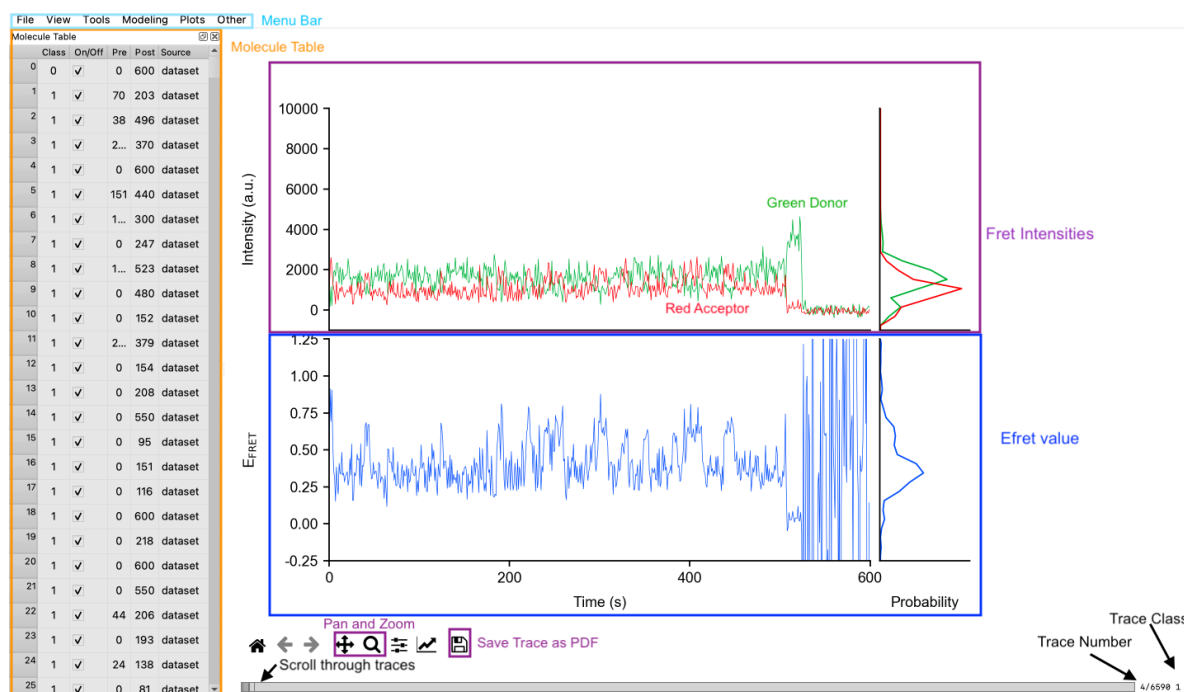

Figure 4.1: labeled UI after loading a dataset

Most actions in tMAVEN do not alter the raw data of a trace, such as gamma and bleed-through corrections. Only cutting traces actually alters the pre and post times which can be seen in the molecule table or on the traces themselves. The user can set the post time by right clicking at the desired point on a graph or left clicking for the pre time. Notice that once a section has been cut out, it appears much fainter on the graph. Reset the pre and post times for a single trace by hitting **R**. Users can also use the square brackets to increase and decrease the post time by 1 point.

7

Other tools include the **pan** and **zoom** features (see Figure). Use the zoom tool to select a section of the graph to enlarge and to allow more precise selection of pre/post time. Additionally, hit **G** to toggle viewing subdivisions on the graph (settings for these can be altered in preferences as described earlier). Another useful tool is **single photobleach detection**, which finds and cuts out photobleaching for only the trace shown. Hit **P** to apply this tool to the shown trace. Photobleach for all traces is also available and is discussed in section 3.5.

But before any manual trace picking is done, it is highly recommended that the user employ the preprocessing tools described in section 3.

5 3. Preprocessing

The tMAVEN tools allow users to skip a lot of the painful preprocessing by culling bad traces, reordering traces by viability, and adjusting the presentation of those traces without altering the raw data, all of which makes data sets easier to process. These features can be found under "tools", and all should be considered before trace picking to optimize the process.

Before going through traces users should navigate to tools/ selection/**order by/cross-correlation** to order traces by cross-correlation since good FRET data corresponds to cross correlated jumps of the donor and acceptor emissions. Notice also the options to **order by** and **Turn on/off** by class. One of the greatest strengths of tMAVEN is the ease of classifying and viewing traces with the molecule table (refer to section 1.2). The power to classify traces greatly enhances use of the algorithms discussed in section 4 by the ease with which one may experiment with inclusion of groups of traces. In this case the user is not forced to go through the whole data set repeatedly.

6 Filter Traces

The **Filter Traces** command, found under tools, allows the user to filter traces by signal to background ratio (SBR). An algorithm that identifies photobleaching (see 3.6) is used to determine the parameters of background. Filter Traces then sorts traces by SBR and allows the user to classify or cull those groups.
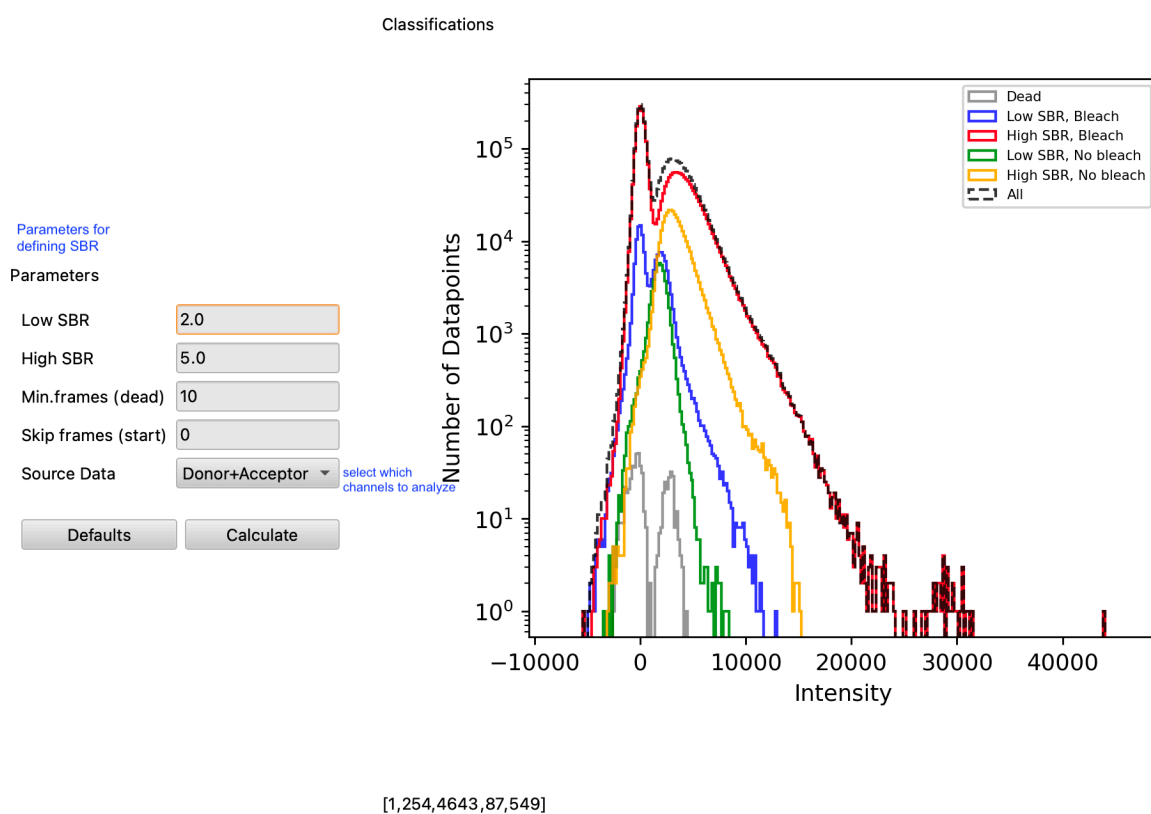
Classifications

Figure 6.1: A graph like this is generated once calculate is hit. The actual culling/classifying is not completed until the user hits process (see figure 3). The graph enables the user a comprehensive view of their data set

Culling in Filter Traces:

Dead	Remove Traces ▼
Low SBR,Bleach	Remove Traces ▼
High SBR,Bleach	Classify 2 ▼
Low SBR,No bleach	Classify 0 ▼
High SBR,No bleach	Nothing ▼

Process

Once calculate is hit and the graph is made, the user can choose what to do with each type of trace (figure 3): ones with low SBR and bleaching, ones with high SBR and bleaching, ones with low SBR and no bleaching, ones with high SBR and no bleaching, and dead traces (traces that bleach before the "minimum frames dead"). Users can repeat the process of filtering and culling until satisfied.

7 Cull

The **Cull** function under tools is relatively intuitive and allows users to discard traces according to class or a maximum/minimum data point value. The latter functions should be used only after photobleach detection (see section 3.6) lest good traces be culled. The data searched for this function includes only Emissions and not E_{FRET} values.

8 Corrections

In this group there are many useful tools such as **bleedthrough**, **gamma**, and **background correction**. As discussed earlier, the actual data is not altered. The program stores a copy of the given data, and the copy is altered by corrections and displayed. For this reason, the corrections should be applied just once for each time a data set is loaded into tMAVEN.

- [Bleedthrough corrections] The default bleedthrough correction is set to 5% and can be adjusted in preferences under `correction.bleedthrough`.
- [Gamma Corrections] The default gamma correction is set to 1 and can be adjusted in preferences under `correction.gamma`.
- [Background] Background is determined for each trace by the last x frames, alterable in preferences under `correction.backgroundframes`, and subtracted from the red emission value.

\end{description}

9 Signal Filters

Also found under tools/corrections are various signal filters that can make traces much easier to read. All smooth the graphs according to different methods, making the E_{FRET} states easier to discern. The filters are found under Tools/Corrections/**Signal Filters**, and can be removed by resetting corrections under corrections/**reset**. Reset corrections before switching between signal filters. The width of the filter window's can also be adjusted in preferences by `correction.filterwidth` which is automatically set to 2. See figure 4 for a look at these filters.

- Gaussian
- Wiener
- Median
- 8-pole Bessel
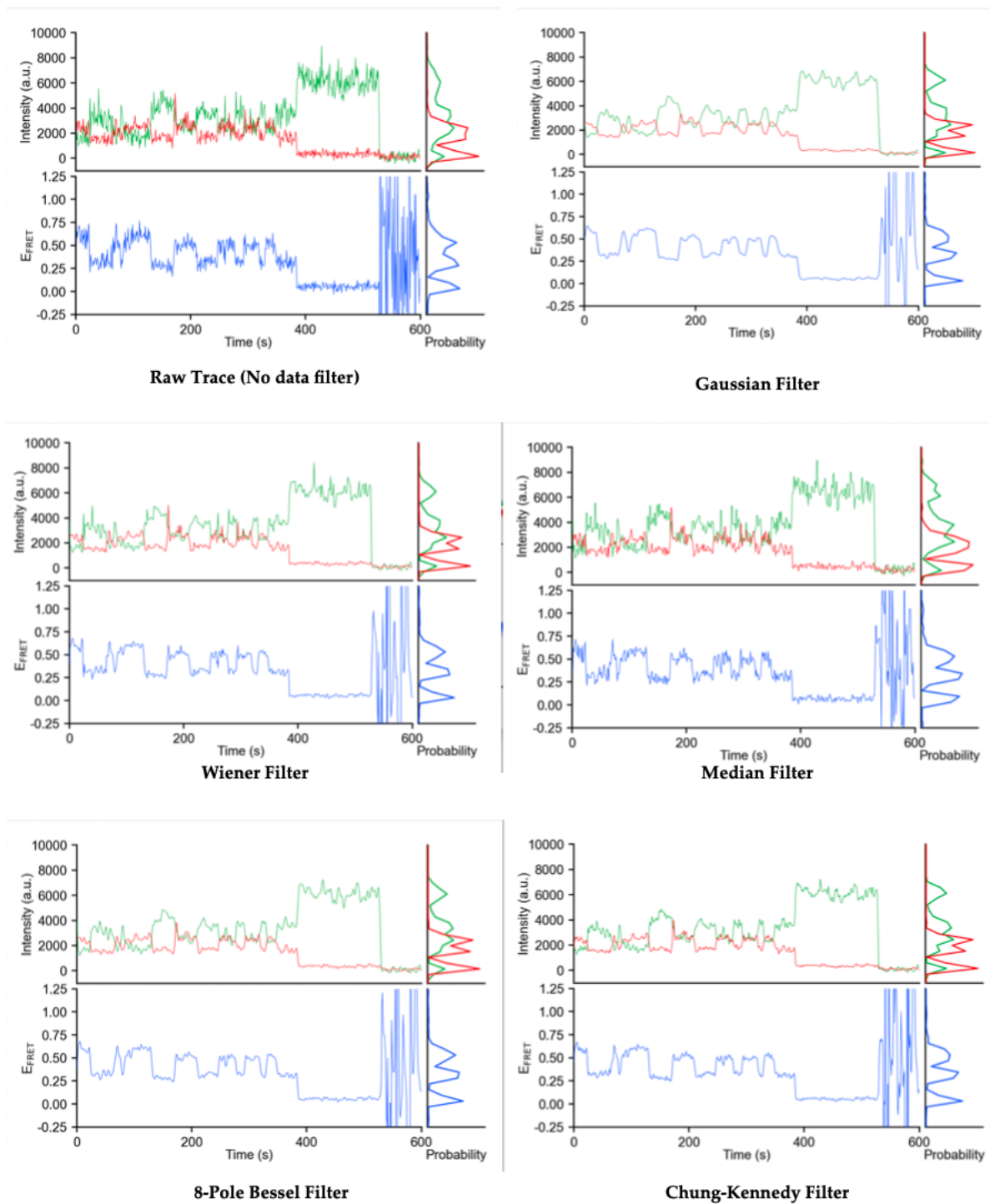- Chung–Kennedy (Chung and Kennedy 1991)

Figure 9.1: The visual effect of the signal filters offered by tMAVEN on the same trace (top left). The benefits of different filters and level of smoothing can be seen here. The Chung-Kennedy filter best maintains the sharp transitions while also reducing noise well. In comparison the Gaussian filter is much more rounded, which makes determining transition times harder.

10 Photobleaching

By navigating to tools/photobleach/**photobleach detection**, the user can run an algorithm that checks each trace for photobleaching and cuts the trace there. Alternatively, users can hit **P** and cut only the trace they are looking at as was discussed in 2.3. The algorithm essentially identifies transitions by maximizing the evidence of a Gaussian between sections of data and placing transitions between those. Photobleaches are then transitions to mean 0 states.

11 4. Modelling

Possibly the most important feature of tMAVEN is its ability to generate models using various algorithms. Users can also use a variety of plots to assess the performance of those models. For more information on Bayesian inference and use of Hidden Markov Models (HMMs) to model single-molecule data (see Kinz-Thompson, Ray, and Gonzalez 2021). A majority of tMAVEN's algorithms use HMMs which model a time series as transitioning between discrete states hidden by noise.
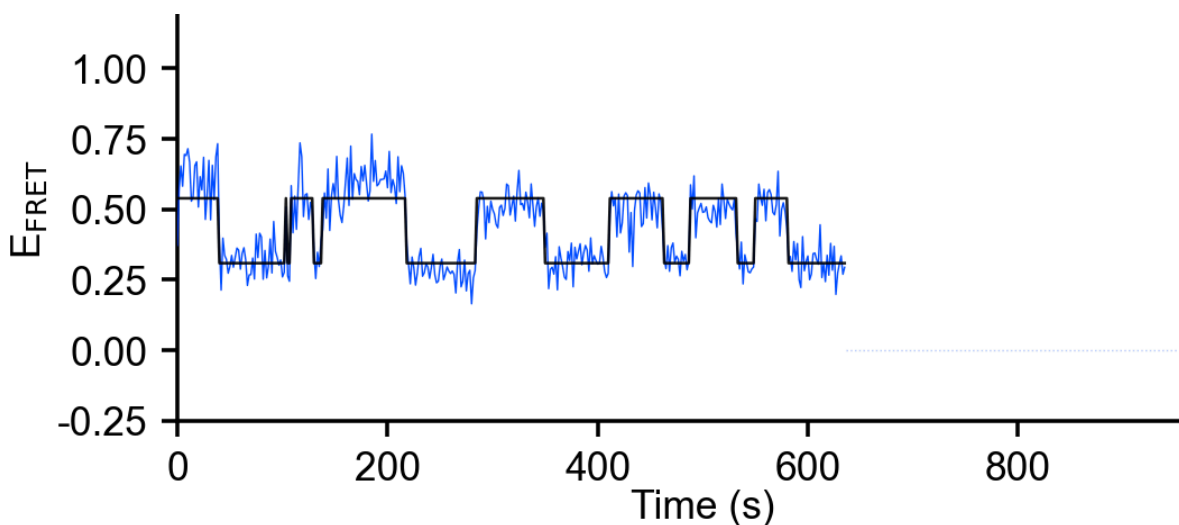

Figure 11.1: After generating a model (see below) tMAVEN shows the idealized trajectory (black line) also called the Viterbi path in black, indicating the most likely state of a molecule at each time point

In running any of these algorithms, it may be useful to also check the log, by navigating to View/**Show Log**. The parameters of a model can be found here, for example: number of determined states, transition matrix, or state means. When generating a model, tMAVEN will only use "on" traces. Viewing the molecule table (1.2) and using the selection tool (2.1) makes toggling molecules on and off much easier.

Additionally, which model is displayed through idealized trajectories may be selected under Modeling/**Change Active**. An individual model can be removed with Modeling/**Clear Active** or all can be removed with Modeling/**Remove Models**. Models can be exported or

imported as hdf5s with Modeling/**Export Active** and Modeling/**Load Active**. If users wish to update the idealized trajectories shown, either after changing pre/post times or including/excluding certain traces, they should click Modeling/**Update idealized**. Note that this will not update the existing model parameters.

#Generating Models There are many models to choose from under Modeling/**FRET Modeling**. One important aspect of these models is whether or not **Model Selection** is utilized. Without model selection, the user must know and input the number of states, however, if that number is unknown, some algorithms may determine it themselves (see Figure). These papers elaborate on that selection: For vbFRET (Bronson et al. 2009) and ebFRET (Meent, Bronson, and Gonzalez 2014)

When running any model, users also have to input some computational settings (see Figure). It is recommended to just leave the defaults. The number of restarts corresponds to the number of times the algorithm will restart with different initializations, in order to ensure this state does not affect the final result. The setting for the convergence determines at what relative change in cost function the algorithm will determine convergence has occurred. In case the convergence doesn't occur, the algorithm will cut off after the max iterations.

Figure 11.2: After selecting a model, tMAVEN prompts users to determine some parameters for the modelling algorithms that inform the prior. In most cases these settings need not be changed aside from number of states. The parameters asked for might change depending on the model type. The algorithms in tMAVEN use a Gaussian gamma function with: a as the shape parameter, b as the rate parameter of the gamma distribution parameters, beta as the Gaussian precision parameter, and alpha as the sole parameter of the Dirichlet distribution (Bishop 2006).

12 Generating a Threshold Model

Under Modeling/FRET Modeling/**Threshold** is the simplest Model. the threshold model assigns every data point to one of two states dependent on whether it is greater than or less than a user-specified threshold. Threshold models calculate average emission values, means, variances, and fractions of the two states visible in the log.

13 Generating a Mixture model

Under Modeling/FRET/**Mixtures** are various models that simply cluster the data from all traces either by **K-means** or by a Gaussian mixture model (GMM) with either a maximum likelihood (**mlGMM**) or a variational Bayesian (**vbGMM**) technique (Bishop 2006). If the number of states is unknown, users can also choose to use model selection (see section 4.2) with variational Bayesian (**vbGMM + Model Selection**). Mixture models calculate average emission values, means, variances, and fractions of k states visible in the log.

14 Generating a Composite HMM Model

Composite models under Modeling/FRET Modeling/**Composite HMMs** create a hidden Markov model (HMM) for each trace and then cluster them in various ways. This single clustered model is then used for the idealized paths shown on each trace.

To generate the HMMs, users can choose either a maximum likelihood algorithm (**mlHMM**, McKinney, Joo, and Ha 2006) or variational Bayesian algorithm (**vbHMM**, Bronson et al. 2009). If the number of states is unknown, users can also choose to use model selection (see Section 4.2).

Once HMMs have been generated, they are clustered in various ways: K-means, vbGMM and Threshold which have been elaborated on in sections 4.3 and 4.4. The means, variances and fractions of the states from the clustered model can be found in the log.

15 Generating Global HMMs

vbConsensus (variational Bayesian, **TMAVENPAPER?**) and **ebHMM** (empirical Bayesian, Meent, Bronson, and Gonzalez 2014) are both consensus methods, in that they generate one HMM from the entire data set. In addition to means, variances, and fractions, the consensus methods yield a transition matrix, which may be found in the log or under **Analyze Dwell Times** as discussed in section 4.7.

16 Dwell Time Plots and Analysis

Once a model is run and the number states and their means and transition matrices are calculated, tMAVEN can analyze a data set in terms of dwell times. Before analyzing, go to modeling/**Analyze Dwell Times**, select your model with **Change Active** and hit calculate (see Figure). To look at a graph for the dwell of a specific state, choose the desired state under rate analysis and input whether it is expected to be a single, double, or triple exponential and hit **run**. Under results, the "Rate type" will read "Dwell Analysis", and rates and coefficients will be shown according to the form of the equation selected. For instance a double exponential will show two rates and two coefficients according to the equation $y = A_1 e^{-r_1 t} + A_2 e^{-r_2 t}$.

Figure 16.1: Before hitting calculate, make sure the intended model is used by looking at "Model Type" in the top left.

Then, either navigate to plots/**Dwell Times** or simply hit **Plot** in the Dwell Times section of Run Dwell Analysis (see Figure). Now that the plot is made (see Figure), a histogram of
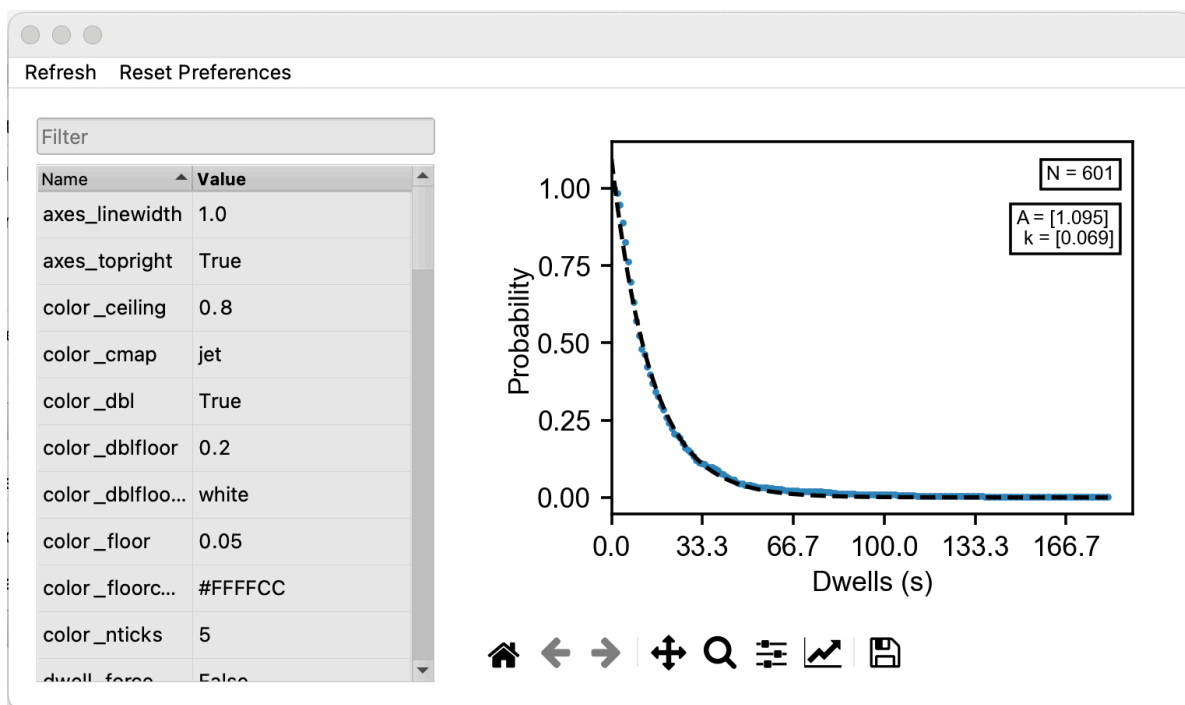
Figure 16.2: A dwell time plot with the data in blue and theoretical dwell graph, from model selected as in figure 7, as a dotted black line. To change the log scale to linear, search on the left for hist_log_y. To toggle the model, search for model_on. To change which state of the model is displayed, search for dwell_state and set to the desired.

the dwell time is displayed, an exponential decay function. The presets show the y axis on a log scale and the model, theoretical dwell, off. Alter these in the preferences on the left (see Figure).

If any series do not appear, or do not update when dwell_state is changed, navigate back to modeling/analyze dwell times and hit calculate again, turn off and on the active model, or hit refresh in the top left corner of the dwell time plot (see Figure).

17 5. Plotting

The histogram feature of tMAVEN is primarily useful for visualization of data and not diagnostics. Navigate to Plots/**FRET hist 1D** to generate a one-dimensional histogram of all E_{fret} emission data, the peaks represent states and should look like K gaussians.
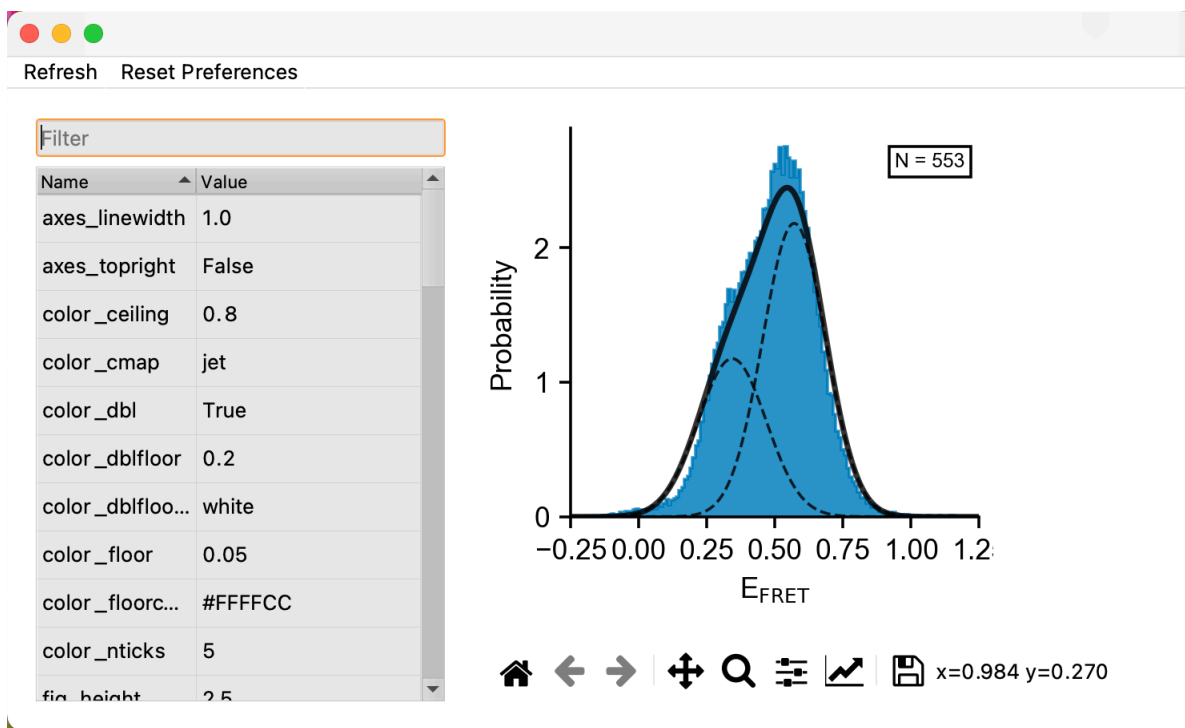

Figure 17.1: A typical one dimensional histogram. Note the number of traces used N in the top right. If a model has been calculated, a population-weighted set of emission distributions is plotted as the solid black with the dotted lines as the individual populations. These lines are not a Gaussian fit of the data, but come directly from the models. Theoretically, better models/processing will result in a better fit of the histogram.

Notable preferences the user might want to change include:

- `hist_force_y` to force the y axis value
- `hist_log_y` puts the y-axis in log scale

- `model_on` will toggle the model as described in Figure
- `hist_false` toggles the histogram itself, blue in the figure
- `hist_color` will change the color of the histogram

The save icon below the graph can be used to export the figure.

18 Two-Dimensional Histograms

The two-dimensional histograms in tMAVEN also have the axis of time. Generate by navigating Plots/**FRET hist 2D** The histogram is automatically smoothed by a median filter which setting can be changed (see below). One of the most useful capabilities is the choice of using post-synchronization. In this case, t=0 represents the time of transition from one specified state to another, see below. Note that post-synch can only operate once a model is run and transitions have been detected.
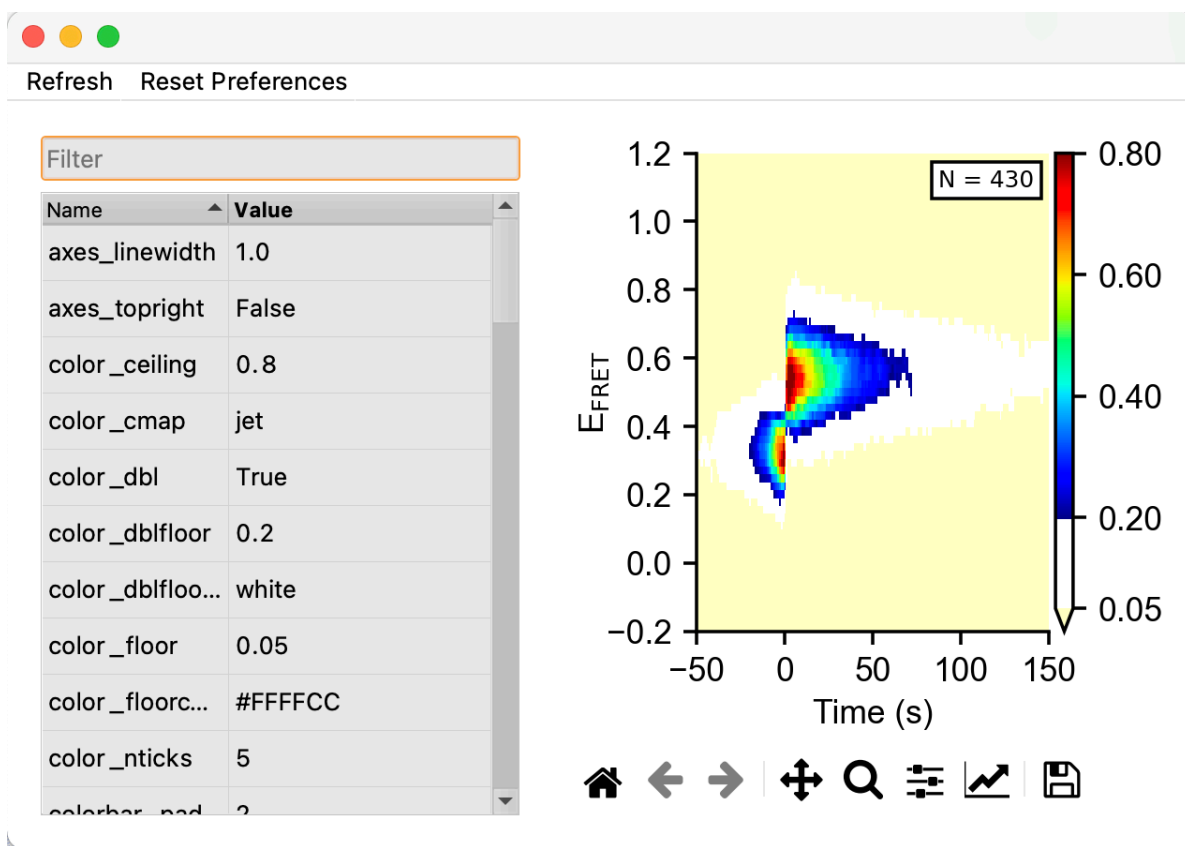

Figure 18.1: A two dimensional histogram with post-sync on using the het color map.

Useful preferences:

- `color_ceiling` sets the ceiling for the color mapping

- `color_dbl` enables two floors
- `color_floor_color` is the lower floor
- `color_cmap` determines the color map, all matplotlib maps are supported
- `hist_smooth_med` toggles the use of a median filter on the graph and `hist_smoothx/y` determine the widths of this filter
- `time_dt` changes the time ticks, set to acquisition time to yield the same axis as traces while `time_nbins` changes the length of time shown.
- `hist_normalizeframe` will normalize the histogram to mitigate photobleaching or additional transition at different times, see Figure

For post synchronization, tMAVEN will use the existing model which has identified transitions so users can essentially monitor all traces after a specific transition in a data set. Notice that N, number of traces used, will likely change since not all traces have all transitions. Small n represents the number of transitions measured not number of traces.

- `sync_postsync` toggles post synchronization
- `sync_singledwell` set to true shows only the dwell before and after transition. In other words, a false setting will also show future transitions.
- `sync_hmmstate_1` shows the pretransition state while `sync_hmmstate_2` shows the post transition state, input "-1" for any other state
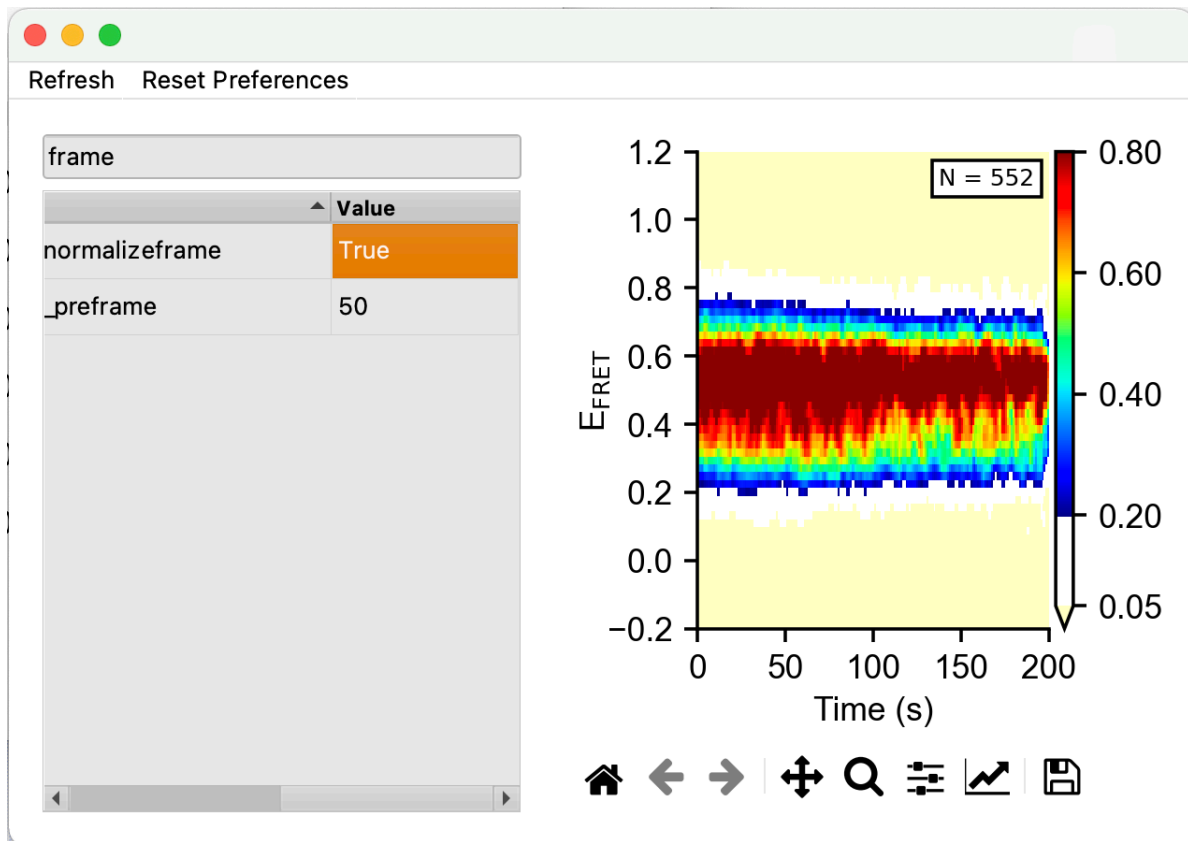
Figure 18.2: The same graph as figure 10 with normalization on.

References

Bishop, Christopher. 2006. *Pattern Recognition and Machine Learning.* Information Science and Statistics. New York: Springer-Verlag. https://www.springer.com/gp/book/9780387310732.

Bronson, J. E., Jingyi Fei, Jake M. Hofman, R. L. Gonzalez Jr, and Chris H. Wiggins. 2009. "Learning Rates and States from Biophysical Time Series: A Bayesian Approach to Model Selection and Single-Molecule FRET Data." *Biophysical Journal* 97 (12): 3196–3205. https://doi.org/10.1016/j.bpj.2009.09.031.

Chung, S. H., and R. A. Kennedy. 1991. "Forward-Backward Non-Linear Filtering Technique for Extracting Small Biological Signals from Noise." *Journal of Neuroscience Methods* 40 (1): 71–86. https://doi.org/10.1016/0165-0270(91)90118-j.

Greenfeld, M., J. W. van de Meent, D. S. Pavlichin, H. Mabuchi, C. H. Wiggins, R. L. Gonzalez Jr, and D. Herschlag. 2015. "Single-Molecule Dataset (SMD): A Generalized Storage Format for Raw and Processed Single-Molecule Data." *BMC Bioinformatics* 16 (3). https://doi.org/10.1186/s12859-014-0429-4.

Kinz-Thompson, C. D., K. K. Ray, and R. L. Gonzalez Jr. 2021. "Bayesian Inference: The Comprehensive Approach to Analyzing Single-Molecule Experiments." *Annual Review of Biophysics* 50 (1): 191–208. https://doi.org/10.1146/annurev-biophys-082120-103921.

McKinney, S. A., C. Joo, and T. Ha. 2006. "Analysis of Single-Molecule FRET Trajectories Using Hidden Markov Modeling." *Biophysical Journal* 91 (5): 1941–51. https://doi.org/10.1529/biophysj.106.082487.

Meent, J. W. van de, C. H. Bronson J. E.and Wiggins, and R. L. Gonzalez Jr. 2014. "Empirical Bayes Methods Enable Advanced Population-Level Analyses of Single-Molecule FRET Experiments." *Biophysical Journal* 106 (6): 1327–37. https://doi.org/10.1016/j.bpj.2013.12.055.